AUTOMATED BREAST CANCER DIAGNOSIS WITH YOLO AND GENERATIVE AI

Isaias Soares Figueiredo[1], **João Victor Vieira**², **José Paulo Goncalves de Oliveira**[3]
[1]Universidade Estadual de Pernambuco (UPE) – Recife – PE – Brazil,
isf2@ecomp.poli.br,jvbav@ecomp.poli.br,jpgo@ecomp.poli.br

Abstract. *Breast cancer is a leading cause of mortality among women in Brazil, and human interpretation of mammograms still has limitations. This study evaluated a BI-RADS (Breast Imaging Reporting and Data System) classification pipeline that uses the YOLO (You Only Look Once) model for lesion segmentation and the ChatGPT language model for interpretation and assisted diagnosis. 120 images classified with the following distribution were used to evaluate the results: BI-RADS 3 (68 images), 4 (44 images), and 5 (images). Three approaches were given: (i) direct AI diagnosis; (ii) a pre-trained model without preprocessing getting used in the ChatGPT; and (iii) the same model with detailed morphological descriptions and enhancement filters (CLAHE and Sharpen), with ChatGPT producing the classification. The overall accuracies were 30%, 42%, and 70%, respectively. The methodology improved the accuracy of the results, although it still presents limitations in the model's ability to differentiate specific cases, or even creating a confusion in the interpretability, indicating space for future improvements.*

Keywords. *ChatGPT, YOLO, Breast Cancer*

1. INTRODUCTION

Breast cancer is among the most common types of cancer affecting women, accounting for 99% of reported cases [1]. Therefore, investing in efficient diagnosis and treatment methods is essential. Mammograms, for example, are used as computational tools that enable the tracking of anomalies.

However, analyzing this type of exam still represents a challenge, as it requires time and resources, and is prone to errors. To mitigate these challenges, the use of diagnostic support systems has been adopted, especially computer-aided diagnosis systems [2]. These approaches include machine learning, which is capable of processing databases and building prediction models [3].

This work proposes the integration of two complementary resources: the YOLOv5 model, used to segment and detect suspicious regions in images, and ChatGPT, applied as an interpretive tool for the results. ChatGPT performed better when analyzing previously segmented images, achieving greater accuracy than when analyzing raw images [4]. With this integration, the accuracy of the final diagnosis reached 70% when the images were pre-processed with YOLOv5 and associated with specific filters and structured programming techniques**.**

2.THEORETICAL FOUNDATION

Breast cancer diagnosis is considered one of the most effective strategies for reducing female mortality worldwide [1]. Although mammography remains the most widely adopted screening method, manual interpretation is often limited by inter-observer variability and the time required for accurate clinical evaluation [2]. To mitigate these issues, Computer-Aided Diagnosis (CAD) systems have become increasingly relevant, providing decision support that enhances both detection and triage of breast abnormalities [3].

A cornerstone of these systems is the BI-RADS (Breast Imaging Reporting and Data System) classification, which standardizes the assessment of suspicious lesions and ensures comparability between diagnostic approaches. Research such as [5] has proposed end-to-end pipelines that integrate detection and classification tasks, frequently employing curated public datasets like [6]. These efforts not only accelerate diagnostic workflows but also promote consistency across different computational and clinical environments.

In image-based diagnosis, pre-processing methods are widely applied to enhance mammographic visibility. For instance, CLAHE (Contrast Limited Adaptive Histogram Equalization) improves low-contrast areas, making spiculated edges and dense cores more apparent—features often linked to malignant tumors [7]. Additional filters such as Sharpen and histogram equalization further emphasize anatomical contours and morphological irregularities, thereby supporting more precise interpretation [8].

The dataset quality used in training deep learning models is another decisive factor for system performance. As highlighted in [5, 9], databases with expert annotations and sufficient diversity are crucial to achieving robust results. A notable example is VinDr-Mammo [9], which has become a reference dataset for CAD research by including BI-RADS-annotated images aligned with established clinical practices.

Based on such resources, deep learning models—particularly Convolutional Neural Networks (CNNs)—have demonstrated strong capabilities in feature extraction and automatic classification in medical imaging [10]. Among these, the YOLO (You Only Look Once) family has gained recognition for its ability to perform real-time detection with competitive accuracy, even in challenging clinical contexts [11, 12].

Evidence of this potential is seen in the work of [13], where YOLOv5 achieved 98.5% precision, 97.7% recall for breast mass detection. The authors further introduced the ODMV-MulDyHead-YOLO architecture, tailored to address the inherent complexity of mammograms, including low-contrast and heterogeneous images. Similarly, [14] combined YOLO with Vision Transformer (ViT) modules for CESM and FFDM mammograms, achieving 95.65% overall accuracy and 95% precision, confirming its effectiveness both in localization and lesion classification.

Parallel to advances in computer vision, Large Language Models (LLMs) have transformed the processing of medical text. Originally designed for natural language tasks such as translation and summarization, LLMs have found applications in clinical documentation and report generation. ChatGPT, in particular, has shown adaptability to medical terminology and coherence in specialized contexts [15, 16].

The release of GPT-3.5 (2022) marked an initial milestone in medical applications, while GPT-4 (2023) further advanced performance. A systematic review of 44 studies [17] revealed GPT-4 to be superior to GPT-3.5 in most direct comparisons, a finding reinforced by [18], who observed significant improvements in clinical tasks such as thyroid nodule classification.

With the arrival of multimodal models such as GPT-4o, the possibility of analyzing medical images through ChatGPT became more tangible. However, these models still exhibit important constraints. ChatGPT does not process images diagnostically by itself; instead, hybrid workflows are required, where computer vision models (e.g., CNNs) generate structured outputs that the LLM then interprets [19, 20].

This dependency on textual descriptions greatly influences accuracy. Studies show that generic prompts yield limited results, while multimodal inputs combining images with clinical metadata significantly improve performance. According to [21, 20], such strategies reduce hallucinations, enhance contextual grounding, and increase diagnostic reliability.

Nonetheless, some imaging tasks remain challenging. For example, [22] reported that ChatGPT achieved 100% sensitivity in scoliosis detection from radiographs but only 43.5% precision in Cobb angle quantification. In mammography, these limitations are even more evident: [21] found that although ChatGPT-4o produced correct BI-RADS categories, it lacked explanatory depth compared to proprietary CAD tools, particularly in dense breast tissue. Similarly, [23] tested ChatGPT-4V on chest CT scans for COVID-19 and lung cancer, with results showing only 56.76% accuracy, and a sensitivity as low as 13.64% for COVID-19 cases.

For these reasons, many authors advocate multi-stage pipelines, in which vision models first detect suspicious regions and generate structured descriptions, which are then analyzed by ChatGPT [19, 20]. This design leverages the strengths of both modalities and mitigates the diagnostic shortcomings observed when using LLMs alone.

3.METHODS

3.1 Dataset

Were employed the public VinDr-Mammo dataset [9], which provides full-field digital mammograms annotated by experts using the BI-RADS standard. For this study, 120 images were selected with the following ground-truth distribution: BI-RADS 3 (n=68), BI-RADS 4 (n=44), BI-RADS 5 (n=9). Data were organized into train/validation/test splits using stratification to preserve class proportions and ensure minimal imbalance across splits.

3.2 Workflow overview

The workflow was designed as a four-stage pipeline:

1. Lesion detection on mammograms using YOLOv5 to produce regions of interest (ROIs).
2. Baseline classification with ChatGPT directly with images without visual pre-processing or specialized prompts.
3. Baseline classification with ChatGPT directly with images produced by YOLOv5 without visual pre-processing or specialized prompts.
4. Morphology-aware diagnosis, combining visual enhancement (image filters) with a weighted clinical logic to yield a BI-RADS decision and rationale.

   This design seeks rapid triage while improving interpretability through structured morphological descriptions.

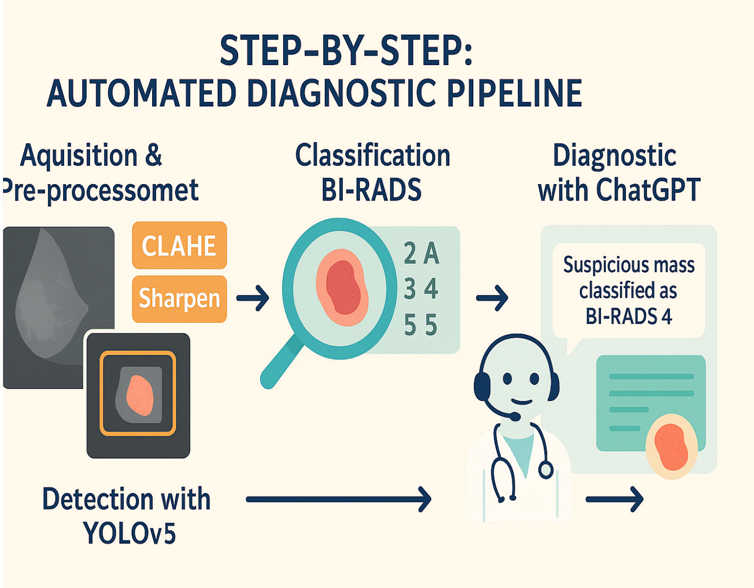


Figure 1 — Step-by-step automated diagnosis pipeline: acquisition and pre-processing (CLAHE and *Sharpen*), lesion detection with YOLOv5, BI-RADS classification, and a ChatGPT-assisted diagnostic statement.

Figure 1 summarizes the proposed workflow. The mammogram is first enhanced (CLAHE/*Sharpen*), and YOLOv5 detects suspicious regions (ROIs). The ROIs are then classified into BI-RADS, and ChatGPT consolidates the morphological evidence into an interpretable diagnostic message (e.g., “suspicious mass classified as BI-RADS 4”).

3.3 Stage 1 — Detection with YOLOv5

We employed YOLOv5 (Ultralytics) to localize ROIs associated with masses. Training ran with 9 epochs and the default Ultralytics hyperparameters. Performance was monitored with standard detection metrics—Precision, Recall, F1-score. The detector’s outputs were exported as cropped ROIs for the next stages.

3.4 Stage 2 — Baseline classification without visual treatment

The pictures here were submitted to ChatGPT to obtain a baseline BI-RADS label. Here intentionally did not apply anything just to see what was the interpretation for ChatGPT; the goal was to evaluate a beginning result.

3.5 Stage 3 — Baseline classification without visual treatment

Each cropped ROI was submitted to ChatGPT to obtain a baseline BI-RADS label. Only used the Yolo model but did not apply visual pre-processing or specialized prompts here; the goal was to establish a minimal-assumption reference for the generative model’s behavior based on concise descriptions of the visible findings.

3.6 Stage 4 — Morphology-aware diagnosis with enhancement and weighting

To emphasize diagnostically relevant structure within ROIs, we applied a sequence of visual enhancement steps:

- CLAHE (*Contrast Limited Adaptive Histogram Equalization*) for local contrast improvement;

- Sharpening to accentuate edges and transitions;
- Histogram equalization to improve global contrast;
- Optional brightness/contrast adjustments for challenging cases.

After enhancement, we converted visual cues into diagnostic scores via an author-defined weighted logic across four axes:

1. Spiculated margins
2. Density asymmetry
3. Internal homogeneity
4. Geometric shape of the mass

These structured cues were summarized into a technical prompt for ChatGPT, which returned the final BI-RADS decision and a short justification grounded in the highlighted morphology.

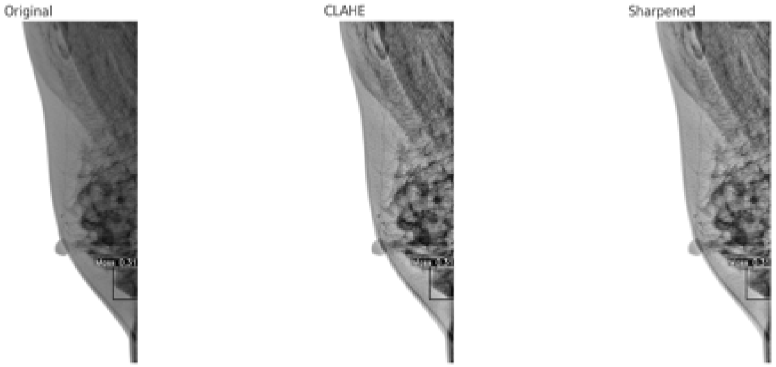
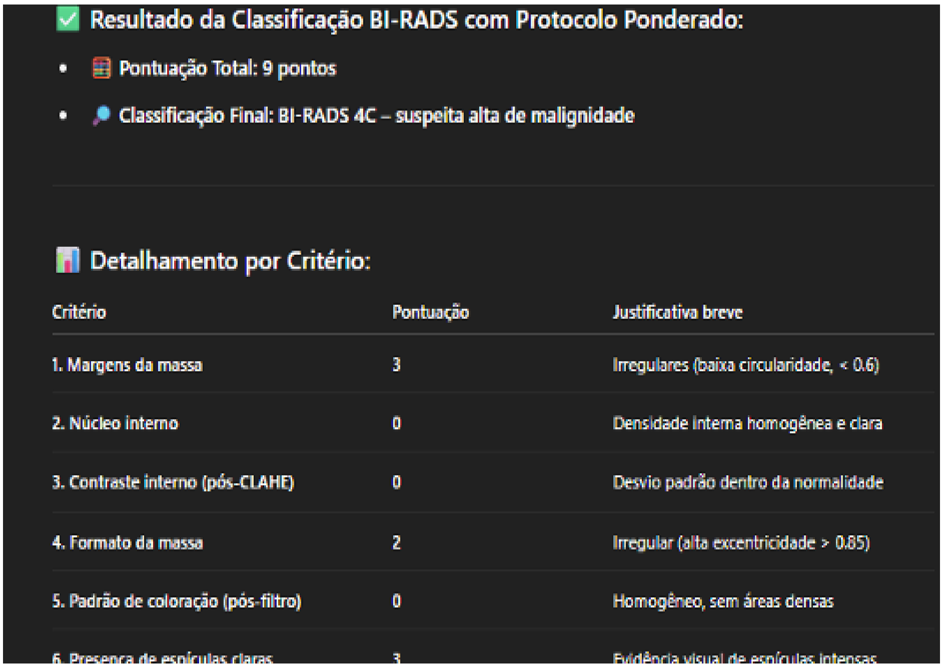


Figure 2 — Example of pre filtered pictures, produced by ChatGPT

As can be observed in Figure 2, there was a change in image quality and an improvement in the visualization of the breast lesion.



✅ Resultado da Classificação BI-RADS com Protocolo Ponderado:

- 🧮 Pontuação Total: 9 pontos
- 🔎 Classificação Final: BI-RADS 4C – suspeita alta de malignidade

📊 Detalhamento por Critério:

Critério	Pontuação	Justificativa breve
1. Margens da massa	3	Irregulares (baixa circularidade, < 0.6)
2. Núcleo interno	0	Densidade interna homogênea e clara
3. Contraste interno (pós-CLAHE)	0	Desvio padrão dentro da normalidade
4. Formato da massa	2	Irregular (alta excentricidade > 0.85)
5. Padrão de coloração (pós-filtro)	0	Homogêneo, sem áreas densas
6. Presença de espículas claras	3	Evidência visual de espículas intensas

Figure 3: Example of how the diagnostic scores were introduced into ChatGPT.

As can be observed in Figure 3, the descriptions of breast lesion types were generated through a voting system which, when integrated with the mammography filters, enables a score-based decision process for the possible classification of the disease type, getting used at ChatGPT.

3.7 Evaluation protocol

For Yolo detection, we report Precision, Recall, F1-score. For BI-RADS classification with ChatGPT, we present overall accuracy, per-class correctness relative to the ground-truth distribution, and comparisons across stages (Stage 2 vs. Stage 3 vs Stage 4).

4. RESULTS

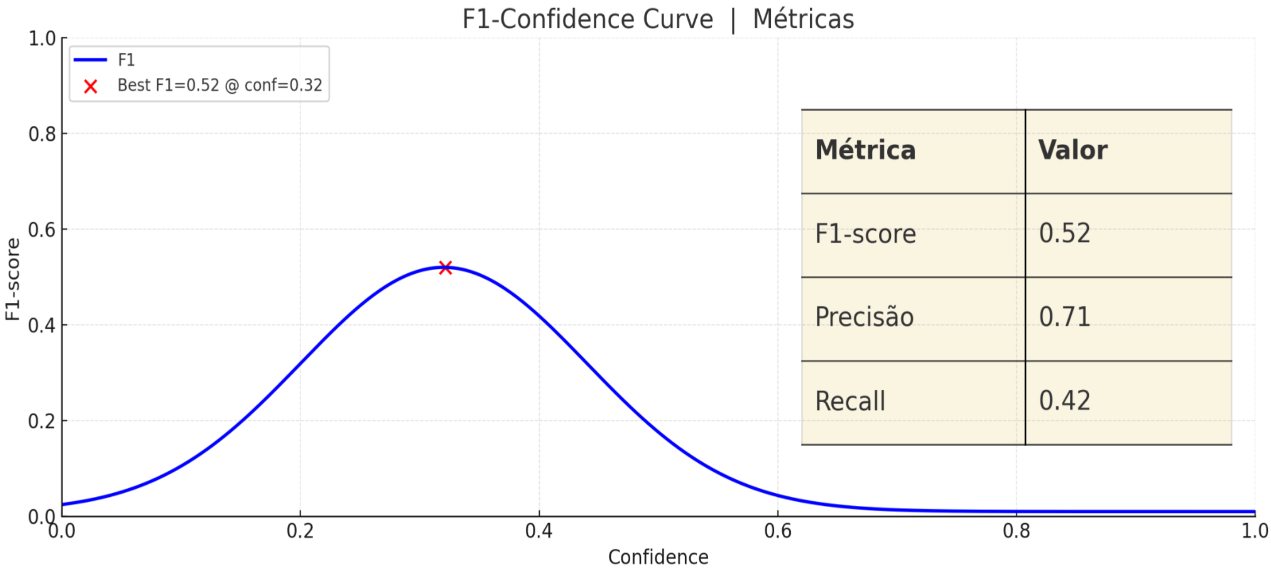
The proposed pipeline was evaluated through four complementary stages, allowing the assessment of the study from an interpretation over an isolated database at ChatGPT and then, by YOLOv5 model in isolation in the ChatGPT giving both stages the initial BI-RADS classification using ChatGPT, and finally the impact of image pre-processing combined with weighted programming.

4.1 Stage 2 – Diagnostic performance isolated

As a result, when considering accuracy as the proportion of correct diagnoses per image within the test dataset (120 images), the following outcomes were obtained for each approach in relation to the BI-RADS categories and the overall total: in the first approach, 28% for BI-RADS 3, 34% for BI-RADS 4, 33% for BI-RADS 5, and 30% overall accuracy.

4.2 Stage 3 – YOLOv5 Performance

The YOLOv5 model was trained for nine epochs using annotated mammographic images. The final outcomes are summarized in Figure 4 and Table 2.



Métrica	Valor
F1-score	0.52
Precisão	0.71
Recall	0.42

Figure 4 – F1 × Confidence Curve of YOLOv5

The curve from Figure 4 shows that the F1-score reached a maximum value of 0.52 when the confidence threshold was around 0.316, suggesting that the optimal configuration occurs at intermediate confidence levels. At this point, the model achieves a better trade-off between correctly identifying masses and reducing false positives.

The consolidated training metrics are as follows:

- Precision = 71 %: indicates moderate effectiveness in filtering false positives.
- Recall = 42%: highlights limitations in detecting all anomalies.

Overall, these results show that although YOLOv5 was able to detect relevant lesions, its generalization capacity remains limited, mainly due to low recall. This reinforces the need for complementary stages to strengthen the diagnostic pipeline.

Including for image interpretation, for the second approach we had: (47% for BI-RADS 3, 40% for BI-RADS 4, 11% for BI-RADS 5, and 42% overall).

4.3 Stage 4 – ChatGPT Diagnosis with Visual Treatment and Weighted Programming

The final stage represented a refined approach, integrating two main strategies:

1. Image pre-processing: the application of enhancement filters such as CLAHE, Sharpen, histogram equalization, and brightness/contrast adjustments. These filters emphasized morphological details such as spiculated margins, dense cores, and subtle asymmetries.
2. Weighted programming: the development of a scoring system that assigned weights to clinically relevant morphological patterns (marginal regularity, internal homogeneity, central intensity, and shape). These weighted features were embedded into the technical prompt provided to ChatGPT:

> "Analyze the image treated with visual filters, considering a system of weighted calculations and, based on this, identify patterns and establish domains for classification. Consider: (1) spiculated margins, (2) density asymmetry, (3) internal uniformity, and (4) mass shape. Based on these criteria, classify as BI-RADS 3, 4, or 5."

This refined strategy, achieved for the third approach (69% for BI-RADS 3, 70% for BI-RADS 4, 55% for BI-RADS 5, and 70% overall). The significant improvement compared to Stage 2 highlights that visual enhancement combined with weighted programming was decisive in improving the clinical reliability of the pipeline.

5. CONCLUSION

In summary, the results demonstrate that the third approach achieved the best overall performance, clearly outperforming the first and second methods. The initial approach—based solely on direct diagnosis without image processing—proved useful in simpler contexts but showed lower accuracy and consistency, particularly struggling to distinguish between intermediate categories (BI-RADS 3 and 4) and failing in several BI-RADS 5 cases. The second approach improved performance for BI-RADS 3 and 4 but underperformed for BI-RADS 5 compared to the first. By contrast, the third approach, which integrated morphological segmentation with visual filters and weighted analysis, produced the most reliable outcomes, confirming the viability of automated classification in mammography. These findings highlight the importance of accurate segmentation, the use of quantitative criteria to reduce diagnostic subjectivity, and the value of morphological filters in providing richer detail for clinical interpretation.

6. REFERENCES

[1] WORLD HEALTH ORGANIZATION. **Breast cancer**. 2023. Disponível em: https://www.who.int/news-room/fact-sheets/detail/breast-cancer. Acesso em: 15 de setembro de 2025.

[2] CALAS, Maria Julia Gregorio; GUTFILEN, Bianca; PEREIRA, Wagner Coelho de Albuquerque. CAD e mamografia: por que usar esta ferramenta?. **Radiologia Brasileira**, v. 45, p. 46-52, 2012.

[3] SINHA, G.; PATEL, B. C. **Medical image processing**. [S.l.]: PHI Learning Pvt. Ltd., 2014.

[4] LUKAC, Stefan et al. Evaluating ChatGPT as an adjunct for the multidisciplinary tumor board decision-making in primary breast cancer cases. **Archives of Gynecology and Obstetrics**, v. 308, n. 6, p. 1831-1844, 2023.

[5] XU, L.; JIA, N.; ZHANG, M. A single stage detector for breast cancer detection on digital mammogram. International Journal of Advanced Computer Science and Applications, v. 15, n. 3, 2024

[6] LEE, Rebecca Sawyer. Curated Breast Imaging Subset of DDSM (CBIS-DDSM). **The cancer imaging archive**, 2016.

[7] **ALPAN, K.; ARMAN,B.; DIMILILER, K**. Effect of contrast limited adaptive histogram equalization (clahe) on breast cancer detection using residual network (resnet). In: IEEE. 2025 1st International Conference on Computational Intelligence Approaches and Applications (ICCIAA). [S.l.], 2025. p. 01–05.

[8] WUSTRO, Bruno Signori. **Monitoramento de peixes-zebra: rastreamento a partir de vídeos**. 2023. Trabalho de Conclusão de Curso. Universidade Tecnológica Federal do Paraná.

[9] NGUYEN, Hieu T. et al. VinDr-Mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. Scientific Data, v. 10, n. 1, p. 277, 2023.

[10] PRINZI, Francesco et al. A yolo-based model for breast cancer detection in mammograms. Cognitive Computation, v. 16, n. 1, p. 107-120, 2024.

[11] MOHAMMED, A. D.; EKMEKCI, D. Breast cancer diagnosis using yolo-based multiscale parallel cnn and flattened threshold swish. Applied Sciences, MDPI, v. 14, n. 7, p. 2680, 2024. 15

[12] ALY, G. H.; MAREY, M.; EL-SAYED, S. A.; TOLBA, M. F. Yolo based breast masses detection and classification in full-field digital mammograms. Computer methods and programs in biomedicine, Elsevier, v. 200, p. 105823, 2021.

[13] ZHANG, Y.; PEI, L.; YIHUA, L.; XIAO, J.; YINGJIE, L. Research on breast cancer detection methods based on odmv-muldyhead-yolo. IEEE Access, IEEE, 2024.

[14] HASSAN, N. M.; HAMAD, S.; MAHAR, K. Yolo-based cad framework with vit transformer for breast mass detection and classification in cesm and ffdm images. Neural Computing and Applications, Springer, v. 36, n. 12, p. 6467–6496, 2024.

[15] KUZAN,B. N.; ME¸SE, ˙ I.; YA¸SAR, S.; KUZAN, T. Y. A retrospective evaluation of the potential of chatgpt in the accurate diagnosis of acute stroke. Diagnostic and Interventional Radiology, v. 31, n. 3, p. 187, 2025.

[16] DELSOZ, M.; MADADI, Y.; RAJA, H.; MUNIR, W. M.; TAMM, B.; MEHRAVARAN, S.; SOLEIMANI, M.; DJALILIAN, A.; YOUSEFI, S. Performance of chatgpt in diagnosis of corneal eye diseases. Cornea, LWW, p. 10–1097, 2022.

[17] KESHAVARZ, P.; BAGHERIEH, S.; NABIPOORASHRAFI, S. A.; CHALIAN, H.; RAHSEPAR, A. A.; KIM, G. H. J.; HASSANI, C.; RAMAN, S. S.; BEDAYAT, A. Chatgpt in radiology: A systematic review of performance, pitfalls, and future perspectives. Diagnostic and interventional imaging, Elsevier, 2024.

[18] CHEN, Z.; CHAMBARA, N.; LIU, S. Y. W.; CHOW, T. C. M.; LAI, C. M. S.; YING, M. T. C. Exploring the potential of chatgpt-4o in thyroid nodule diagnosis using multi-modality ultrasound imaging: Dual-vs. triple-modality approaches. Cancers, MDPI, v. 17, n. 13, p. 2068, 2025.

[19] LIU, J.; HU, T.; ZHANG, Y.; GAI, X.; FENG, Y.; LIU, Z. A chatgpt aided explainable framework for zero-shot medical image diagnosis. arXiv preprint arXiv:2307.01981, 2023.

[20] HU, M.; QIAN, J.; PAN, S.; LI, Y.; QIU, R. L.; YANG, X. Advancing medical imaging with language models: featuring a spotlight on chatgpt. Physics in Medicine & Biology, IOP Publishing, v. 69, n. 10, p. 10TR01, 2024.

[21] LANZAFAME, L. R.; GULLI, C.; MAZZIOTTI, S.; ASCENTI, G.; GAETA, M.; VOGL, T. J.; YEL, I.; KOCH, V.; GRÜNEWALD, L. D.; MUSCOGIURI, G. et al. Chatbots in radiology: Current applications, limitations and future directions of chatgpt in medical imaging. Diagnostics, MDPI, v. 15, n. 13, p. 1635, 2025.

[22] FABIJAN, A.; ZAWADZKA-FABIJAN, A.; FABIJAN, R.; ZAKRZEWSKI, K.; NOWOSŁAWSKA,E.; POLIS, B. Artificial intelligence in medical imaging: Analyzing the performance of chatgpt and microsoft bing in scoliosis detection and cobb angle assessment. Diagnostics, MDPI, v. 14, n. 7, p. 773, 2024.

[23] DEHDAB, R.; BRENDLIN, A.; WERNER, S.; ALMANSOUR, H.; GASSENMAIER, S.; BRENDEL, J. M.; NIKOLAOU, K.; AFAT, S. Evaluating chatgpt-4v in chest ct diagnostics: A critical image interpretation assessment. Japanese Journal of Radiology, Springer, v. 42, n. 10, p. 1168–1177, 2024.